THE PERCEPTUAL EFFECTS OF ANTICIPATORY LABIAL ACTIVITY IN FRENCH

*+Rudolph Sock, *Véronique Hecker and *Marie-Agnès Cathiard
*Institut de Phonétique de Strasbourg - Université Marc Bloch
+Institut de la Communication Parlée de Grenoble - Université Stendhal
*Laboratoire de Psychobiologie du Comportement Moteur et des Sports - Université Marc Bloch

ABSTRACT
This research examines the relationship between an anticipatory labial gesture and the perception of an upcoming rounded vowel. More specifically, the contribution of anticipatory coarticulation to the perception of the rounded vowel is examined in both the motoric and acoustic levels. The robustness of the temporal extent of the perceptual effect is evaluated under increased speaking rate. The paradigm consists in generating speech samples by a representative French speaker, then segments are "gated-out" and listeners are asked to judge what the "gated-out" segments were. The issues of acoustic and visual perception of anticipatory vocalic gestures are finally discussed, in trying to provide a general understanding of anticipatory phenomena in speech.

1. INTRODUCTION
Anticipatory gestures — understood here as the spreading or the extension of certain gestures to adjacent or neighbouring segments — are usually regarded as being of utmost importance in fluent speech. It is also known that listeners exploit early cues related to upcoming speech elements. This has been demonstrated in the domain of visual perception for French [1]. In the domain of speech perception, on the auditory-acoustic level, known results [2] do not take into account the relationship between the articulatory level and its acoustic efficiencies. This research, closely following that carried out for Swedish [3], thus presents data regarding three precise questions: (1) Does the anticipatory protrusion gesture contribute to the auditory perception of the French rounded vowel [y]; (2) If such is the case, what is the temporal extent of the perceptual effect; (3) How does speech rate affect the perceptual extension of this anticipatory gesture?

2. METHODS
2.1. Speaker and speech sample
The speaker was an adult male native speaker of French (Jean-Luc). He had no speech or hearing defects. The speech sample consisted of the sentence "Ces deux Sixtes sculptèrent" that provides the core sequenceV1CnV2, where V1 is the spread vowel [i], Cn is a string of 5 non-labial consonants, and V2 is the rounded vowel [y]. This is one of the sentences used in the elaboration of the Movement Expansion Model-MEM [4, 10]. The speaker produced the sentence several times at a normal conversational rate, and at a self-selected fast rate.

2.2. Acoustic and movement data
A simultaneous face and profile video recording of the speaker was obtained, in synchronous with the audio, using the Institut de la Communication Parlée Speech-Video Station [5]. 50 frames per second, in black and white, were thus acquired. Only a representative example of the acoustic signal — one without an intrusive schwa in the consonant string — and the corresponding video signal were retained for each speech rate. The digitized acoustic signals were then visualized synchronous with the result of the processed video signals, using an adapted speech editor. Two acoustic events (illustrated by the arrows on Figure 1, and corresponding to the onset and offset of a clear formant structure for vowels [i] and [y] respectively), served to determine the temporal extent of the consonant string. Among the different dynamic parameters extracted from the video signal, only the protrusion movement of the upper lip was retained to determine the anticipatory timing of the lip rounding gesture, with respect to the acoustic onset of the subsequent rounded vowel, and the related perceptual effects.

Previous research studies have shown that the upper lip protrusion movement is indeed a good candidate in reflecting the vowel rounding gesture [4, 5]. The kinematic events, detected on the first and second derivatives of the upper lip position signal, are shown in Figure 1: movement onset or Minimum Protrusion (filled circle); Peak Acceleration (empty triangle); Peak Velocity (filled triangle), and Maximum Protrusion (empty circle). The purpose of analyzing the position signal was to be able to determine the occurrence of each kinematic event with respect to the acoustic onset of the rounded vowel, thus having motoric events as landmarks for the gating decisions made on the acoustic signals.

2.3. Gating procedures and test tape elaboration
Gaitings were made on the two digitally stored acoustic signals, one for the normal speech rate and the other for the fast rate. The acoustic obstruent interval was compressed from 394 ms in normal speech to 326 ms in fast speech; so also was Movement Time (which is the distance between Minimum Protrusion and Peak Protrusion), from 239 ms in normal speech to 160 ms in fast speech, thus confirming an adequate execution of the speech rate task required of the speaker. The five gaitings, located at different temporal distances from the acoustic onset of the vowel were the following: Gate 1, based on Minimum Protrusion was at 280 ms in normal speech and at 194 ms in fast speech; Gate 2, based on Peak Acceleration was at 200 ms in normal speech and at 134 ms in fast speech; Gate 3, based on Peak Velocity, was at 80 ms in normal speech and at 74 ms in fast speech; Gate 4, based on Maximum Protrusion, was at 41 ms in normal speech and at 34 ms in fast speech; Gate 5
corresponded, in both speech rates, to the onset of the clear formant structure of the vowel. Thus each "gated" sentence consisted of the sequence [sedɔs]... plus an increasing amount of the acoustic information from the consonant string [kitik] preceding the rounded vowel. The test tape contained 10 gated out stimuli (5 in each speech rate) that were then randomized twice, one series in normal speech and the other in fast speech, resulting in two files containing each 10 stimuli. A bip, serving to alert the listeners as to the imminence of a stimulus, preceded each utterance by 1.4 seconds. The interstimuli interval was 4 seconds with a 10 second pause following each tenth item. The test tape began with a training list of 10 items, then continued with the 20 stimuli.

2.4. Listener judgements

Twenty-two adults, all native speakers of French, served as listeners for the perceptual experiment. They were all naive with regards to the purpose of the tests, and had normal speech and hearing. The tests were carried out in a sound-treated room at the Phonetics Institut of Strasbourg, designed for group listening experiments. The subjects were told that they were going to hear one of the following three truncated French sentences: (1) "Ces deux Sixtes skïerent"; (2) "Ces deux Sixtes scuptïrent"; (3) "Ces deux Sixtes scapïrent". However, in this particular experiment, only the target sentence n° 2 was on the test tape, the other two only serving as distractors. Nevertheless, results from a similar experiment show that when these two sentences were effectively included in the test tape, no significant change was observed in the results [7]. The test tape was played simultaneously to all twenty-two subjects. Score sheets were provided, and during the 4 second interval between test stimuli, the subjects had to perform two tasks for each of the 20 utterances: (1) mark with a cross which of the three vowels [i, a, y] he or she believed to have been "gated out" of the particular test stimuli, and (2) check on a 5-point scale the level of confidence with which they were making the confident ratings dropped off progressively as the gatings approach the target vowel, i.e. 80 ms before the acoustic onset of the rounded vowel. The perceptual effect of this anticipatory movement is maintained in fast speech, as 85% of correct identification is obtained at this gating, located at 74 ms before the vowel. Moreover, 71% of correct identification was already possible in fast speech at the previous gating, which corresponds to Peak Acceleration of the upper lip protrusion movement, situated at 134 ms before the acoustic onset of the vowel. Let us recall that although the latter identification scores were backed by a "2" point confidence rating, identification scores proved to be around 95% on average. Nevertheless, the latter result should be taken with caution; this point will be discussed later. At Maximum Protrusion, located at 41 ms before the vowel in normal speech, scores were at 93% of correct identification, and in fast speech, where Maximum Protrusion was located at 34 ms from the onset of the vowel, correct identification was superior to 92%. Identification was absolute (100%) in both speech rates at the last gating, corresponding to the onset of the clear formant structure for the rounded vowel.

3. RESULTS

The mean level of confidence rating, averaged over all 20 utterances by all 22 listeners (totalling 440 responses) was 3 with a standard deviation of 1, thus showing that overall judgements were made with a good degree of confidence, in both speech rates. Moreover, the percentage of correct responses was highly correlated with confidence ratings (r=0.88 in normal speech and r=0.74 in fast speech), indicating that while being confident subjects were at the same time being highly performant, with regards to the identification task required of them (see Figure 2). Also, subjects seemed to be quite severe with the use of confidence ratings on the subjective 5-point scale, as confidence points "2'" and "3" revealed high correct identification scores in both speech rates, with the curve attaining ceiling values between points "3" and "5". Further, as shown in Figure 3, the confident ratings drop off progressively as the gatings approach the target vowel, i.e. as the amount of available sensory information decreases (r=0.91 and r=0.92 in normal and fast speech respectively). Nevertheless, the suitable model is not that of a linear regression but of a hyperbola, with decrease in confidence ratings being less sensitive as gatings are temporally distant from the onset of the vowel. When very unsure as to the identity of the gated out vowel, listeners largely used the vowel [i] as a "waste-basket" response, since the percentage of [i] responses remarkably exceeds that of the [a] responses at the most extreme gating (Gate 1). This is explained elsewhere [8], less in terms of proximity between [i] and [y] in the acoustic space, than with regards to listeners reacting to the high frequency spectrum of the fricative noise of the second [s] in the word "Sixte", dominant in the vicinity of that extreme gating and perceptually close to [i], that has a relatively high frequency formant structure.

Results given in Figure 4 are in line with what has previously been reported in the literature, as to listeners being able to guess the identity of vowels when those vowels had been gated-out of the acoustic signal [2, 3]. The percent correct identification is high very close to the vowel and decreases as gating moves away from this target vowel (r=0.98 and r=0.95 in normal and fast speech respectively). In normal speech, 79% of correct identification occurs at Peak Velocity, associated with the upper lip protrusion movement, i.e. 80 ms before the acoustic onset of the rounded vowel. The perceptual effect of this anticipatory movement is maintained in fast speech, as 85% of correct identification is obtained at this gating, located at 74 ms before the vowel. Moreover, 71% of correct identification was already possible in fast speech at the previous gating, which corresponds to Peak Acceleration of the upper lip protrusion movement, situated at 134 ms before the acoustic onset of the vowel. Let us recall that although the latter identification scores were backed by a "2" point confidence rating, identification scores proved to be around 95% on average. Nevertheless, the latter result should be taken with caution; this point will be discussed later. At Maximum Protrusion, located at 41 ms before the vowel in normal speech, scores were at 93% of correct identification, and in fast speech, where Maximum Protrusion was located at 34 ms from the onset of the vowel, correct identification was superior to 92%. Identification was absolute (100%) in both speech rates at the last gating, corresponding to the onset of the clear formant structure for the rounded vowel.

4. SUMMARY AND GENERAL DISCUSSION

Taken together, these results (identification scores and confidence ratings) tell us that neither Minimum Protrusion (or the onset of the movement), nor Peak Acceleration, which were the first two gatings, seem to have any direct anticipatory perceptual effect. In other terms, neither the onset of the protrusion movement, nor the kinematic attribute of the movement, Maximum Acceleration, are robust components of the audible anticipatory labial gesture. The gesture becomes audible as from Peak Velocity (80 ms and 74 ms before the onset of the vowel in normal and fast speech respectively), when high identification scores (82% on average) are backed by strong confidence ratings in both speech rates. It is thus the kinematic span Peak Velocity-Maximum Protrusion that seems to constitute the anticipatory acoustic efficient portion of the entire protrusion gesture (without considering the efficient carryover façade during unrounding for the subsequent [I]).
Despite the predominant high frequency energy concentration, characteristic of the voiceless consonant string, the subsequent rounded vowel becomes highly audible as soon as the labial gesture attains its peak velocity. Increasing speech rate tends to enhance — in relative terms — the temporal extent of the perceptual effect. This finding seems to be a tendency only, as stated earlier, high identification scores occuring at 134 ms before the acoustic onset of the vowel in fast speech are not systematically backed by high confidence ratings. In fact, results observed show, in general (see regression lines and curves in Figures 2 to 4), that the perceptual effects of anticipatory coarticulation are structurally the same, in the two speech rates, without systematic significant differences between mean data points.

That Peak Velocity corresponds to the beginning of the robust efficient portion of the protrusion vocalic gesture, seems to support previous findings that reveal remarkable regularities between kinematic phasing patterns, based on timing relations between peak velocities, and acoustic phasing patterns, portraying the timing between onsets and offsets of clear vocalic formant structures [9]. In the present experiment, although increasing speech rate did provoke a compression of Movement Time, or the timing of the obstruct consonants (from 239 ms to 160 ms), the occurrence of Peak Velocity, relative to the rounded vowel, was not perturbed (80 ms vs. 74 ms in normal and fast speech respectively), presumably explaining similarities in the temporal extent of the anticipatory perceptual effect in the two speech rates. However, according to reported data for French and to the predictions of the Movement Expansion Model [4, 10], the protrusion Movement Time may expand as a function of the number of consonants in the obstruct interval, and of the speaker. This elasticity of the movement time is potentially capable of modifying the occurrence of peak velocity, relative to the acoustic onset of the vowel. If such were the case, we posit that significant differences in peak velocity occurrences would provoke noticeable differences in coarticulatory perceptual extents. In fact, results — based on acoustic data only — obtained in a related experiment from two speakers producing the same "gated-out" sentences [8], show within a speaker, a similar global behaviour in the anticipatory perceptual extent of listeners, and a different global pattern for listeners, from one speaker to the other. Given that experimental artefacts and other factors, such as differences in speaker intelligibility or listener auditory capabilities, had been proven irrelevant, we speculate that these global differences in listener behaviour are largely due to configurational differences in the façade of the protrusion gesture, and particularly to the audible portion of that gesture, being robust as from Peak Velocity. These conjectures are presently being verified experimentally.

In the visual domain [1], the anticipatory gesture, in unimodal perception, may be identified 160 ms away from the acoustic onset of the vowel, with an identification score of 90%. This is a demonstration of the superiority of visual information on auditory information, with regards to coarticulation, since such an anticipatory extent is never attained in the acoustic domain. However, like in the acoustic domain, the extent of the visual anticipatory gesture depends largely on the profile of the articulatory signal. In bimodal perception, it has been shown that desynchronizing the coherence of the audio-visual flux [10], by reducing the duration of the natural lead of visual anticipation on the acoustic information, may perturb correct identification of the vowel, when the onset of the sound coincides with the visual boundaries.

Our conclusions are thus the following: in speech the anticipatory protrusion gesture is perceptually an efficient component of the production of a rounded vowel (especially when this feature is linguistically relevant, like in French); its temporal extent is greater in the visual domain (situated at Maximum Acceleration of the gesture) than in the acoustic domain. These visible and audible events are not synchronous since they reflect articulatory-acoustic non-linearities, which seem to favour perceptual coherence, as highlighted by desynchronization tests. We believe, however, that speech is most of all a contribution of these dynamic factors to vocal tract and facial configurational constraints (recovering shape from motion), tightly coupled with the acoustic level.

NOTE

1. This research is dedicated to the memory of Christian Benoît.

REFERENCES


Figure 1. An example of the audio and the upper lip protrusion signal for the French Sentence "Ces deux Sixtes sculptèrent". (Abry and Lallouache, 1995).

Figure 2. Mean values of percent correct responses.

Figure 3. Mean values of confidence ratings at different gatings in relation to confidence level.

Figure 4. Mean values of percent correct responses at different gatings.